CY 7790

Special Topics in Security and Privacy: Machine Learning Security and Privacy
Fall 2021

Alina Oprea
Associate Professor
Khoury College of Computer Science

September 13 2021

Introduction

- Ph.D. at CMU, 2007
  - Research in applied cryptography, data security, and cryptographic file systems
- RSA Laboratories, 2007-2016
  - Cloud and storage security, applied cryptography, game theory for security
  - ML/AI in security
- Northeastern Khoury College – since Fall 2016
  - NDS2 Lab part of the Cybersecurity and Privacy Institute
  - Machine learning for security applications: attack detection, IoT, connected cars, collaborative defenses
  - Adversarial machine learning: study the vulnerabilities of ML in face of attacks and design defenses
  - Privacy in machine learning: auditing, memorization

TA Introduction

- Giorgio Severi
  - 4th year PhD student at Northeastern
  - Working on adversarial ML for cyber security applications
  - Part of the NDS2 research lab

Class Introduction

- Enrollment of 18

- Research area (if PhD student)
- What topics you are interested in adversarial ML
- Something we cannot read online about you!

CY 7790 Course objectives

- Provide in-depth coverage of adversarial attacks on ML:
  - Evasion attacks at inference time
  - Poisoning attacks at training time
  - Privacy attacks
- Learn how to classify the attacks according to the adversarial objective, knowledge, and capability
- Discuss adversarial attacks in real-world applications: cyber security, NLP, etc.
- Understand existing methods for training robust models and the challenges of achieving both robustness and accuracy
- Discuss fairness issues in machine learning that might exacerbate existing risks of adversarial attacks
- Read and discuss research papers in adversarial ML as a group
- Work on a research project in a team

Course Information

- Website:
  www.ccs.neu.edu/home/alina/classes/Fall2021
- Gradescope: gradescope.com
- Communication: piazza.com
- E-mail:
  - Alina: a.oprea@northeastern.edu
  - Giorgio: severi.g@northeastern.edu

Class Outline

- Introduction – 2 weeks
  - Review of machine learning and deep learning
  - Taxonomy of adversarial ML
- Evasion attacks and defenses - 2 weeks
  - White-box, black-box attacks
  - Adversarial training and certified defenses
- Poisoning attacks – 2 weeks
  - Availability, targeted, backdoor, federated learning
- Application domains – 1 week
- Privacy attacks and defenses: 2 weeks
  - Membership inference, memorization
  - Differential privacy, auditing
- Fairness of AI – 1 lecture

Policies

- Instructors
  - Alina Oprea
  - TA: Giorgio Severi
- Schedule
  - Mon and Thu 11:45am – 1:25pm EST
  - Office hours:
    - Alina: Thursday 4:00 – 5:00 pm
    - Giorgio: Monday, 4:00 - 5:00 pm
- Online resources
  - Use Piazza for questions and discussion
  - Gradescope for paper summaries and assignments

Grading

- Assignments – 10%
  - 2 assignments (one on ML and one on adversarial attacks)
- Paper summaries – 10%
  - Read and submit paper summaries before every class
- Discussion leading – 15%
  - Lead discussion in several classes (team of 2-3 students)
- Scribing – 15%
  - Write notes for 2 lectures
- Final project – 50%
  - Select your own project topic related to robustness, privacy, or fairness of AI (teams of 2)
  - Two types of projects: research or SoK
  - Project proposal, milestone, presentation at end of class, and written report

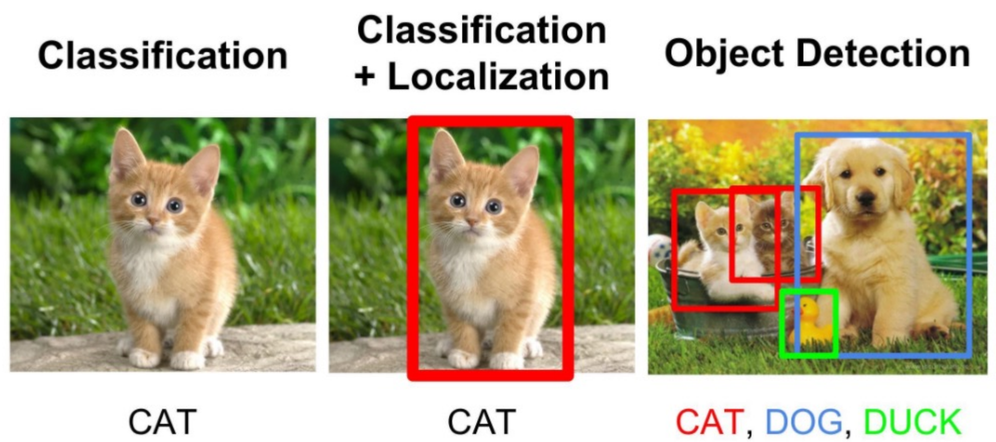
Academic Integrity

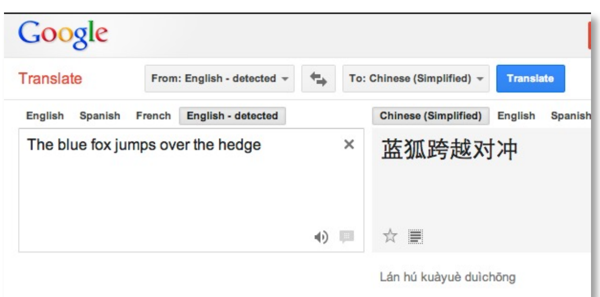
- Homework / paper summaries are done individually
- Class project is done in teams
- Rules
  - Can discuss with colleagues or instructors
  - Can post and answer questions on Piazza
  - Code cannot be shared with colleagues
  - Cannot use code from the Internet
    - Use python or R packages, but not directly code for ML analysis written by someone else
- NO CHEATHING WILL BE TOLERATED!
- http://www.northeastern.edu/osccr/academic-integrity-policy/

ML Resources

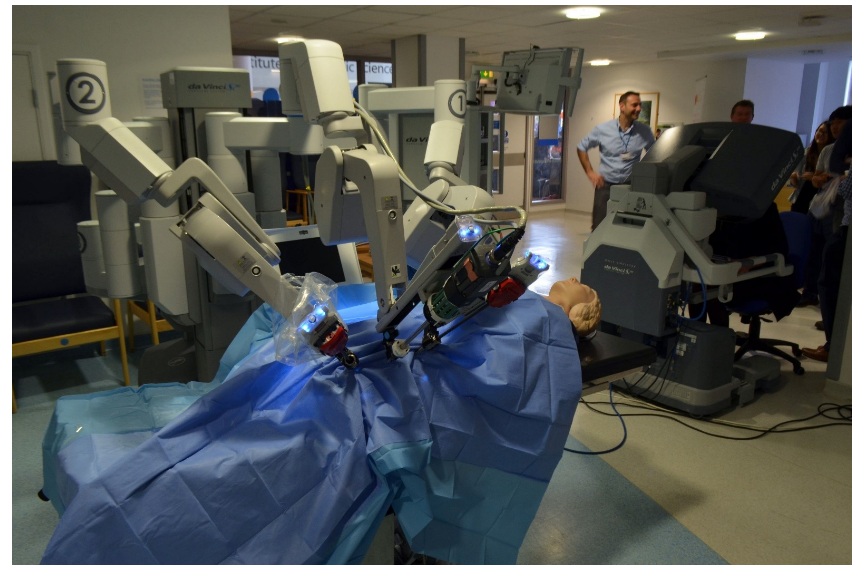
- Trevor Hastie, Rob Tibshirani, and Jerry Friedman, Elements of Statistical Learning, Second Edition, Springer, 2009.
- Christopher Bishop. Pattern Recognition and Machine Learning. Springer, 2006.
- A. Zhang, Z. Lipton, and A. Smola. Dive into Deep Learning
- Lecture notes by Andrew Ng from Stanford
- DS 4400 lecture notes: http://www.ccs.neu.edu/home/alina/classes/Spring2021/
- Trustworthy ML paper list: https://trustworthy-machine-learning.github.io/

Today's Applications of AI





Fast Forward in the Near Future



- More uses in critical applications (smart cities, medicine)

What is Your Favorite ML / Deep Learning Application?

Applications of ML

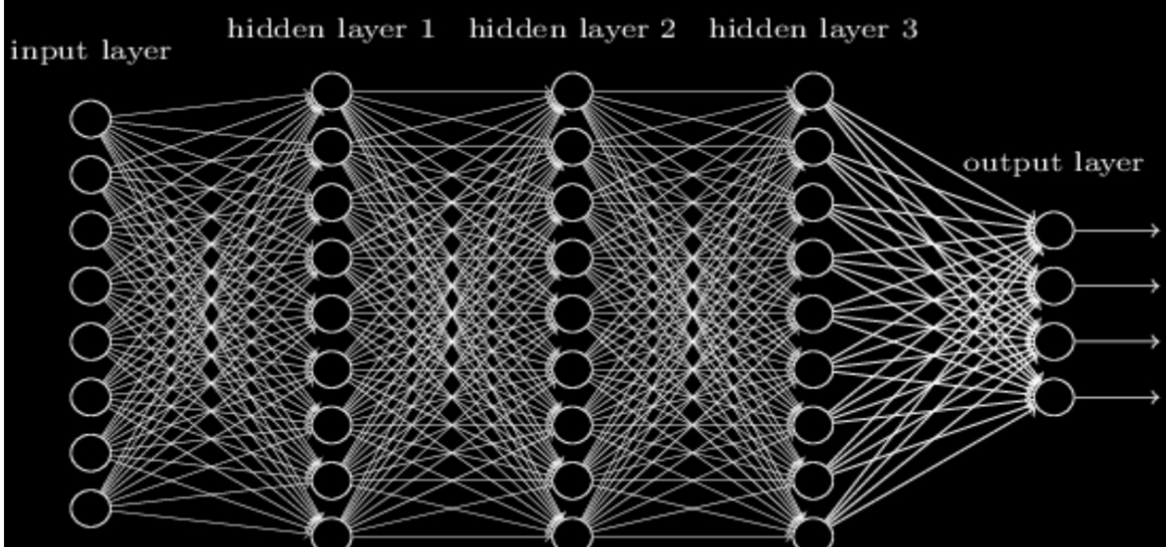
- Healthcare
- Vision
- NLP
- Speech recognition
- Self-driving cars
- Stock market analysis
- Recommendations
- Sentiment analysis
- Human behavior
- Quality of life

- Business
- Sports
- Bots / chatbots
- Science / engineering
- Bioinformatics
- Precision medicine
- Unsupervised learning
- Reinforcement learning

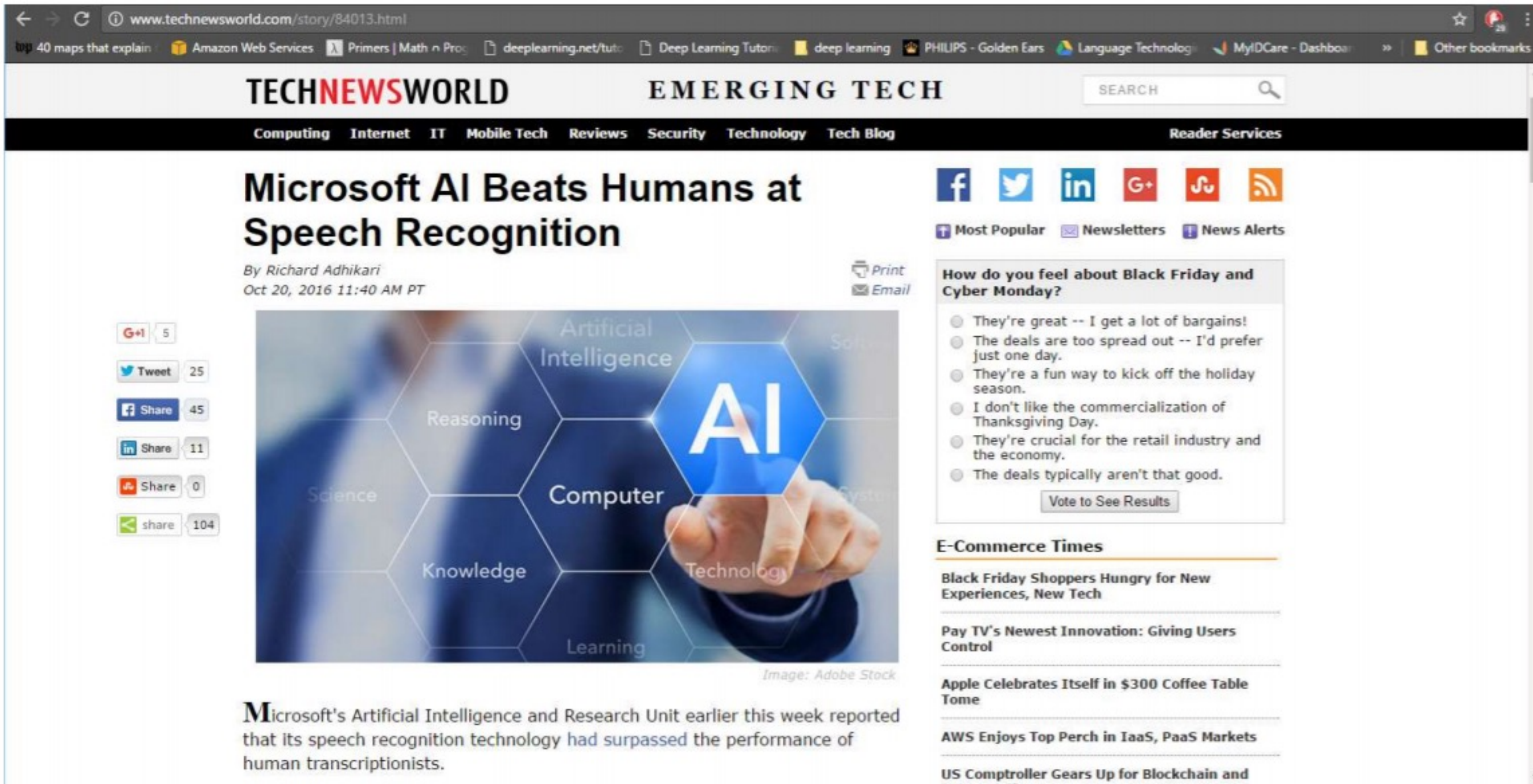
Deep Learning

Neural networks return and excel at image recognition, speech recognition, …

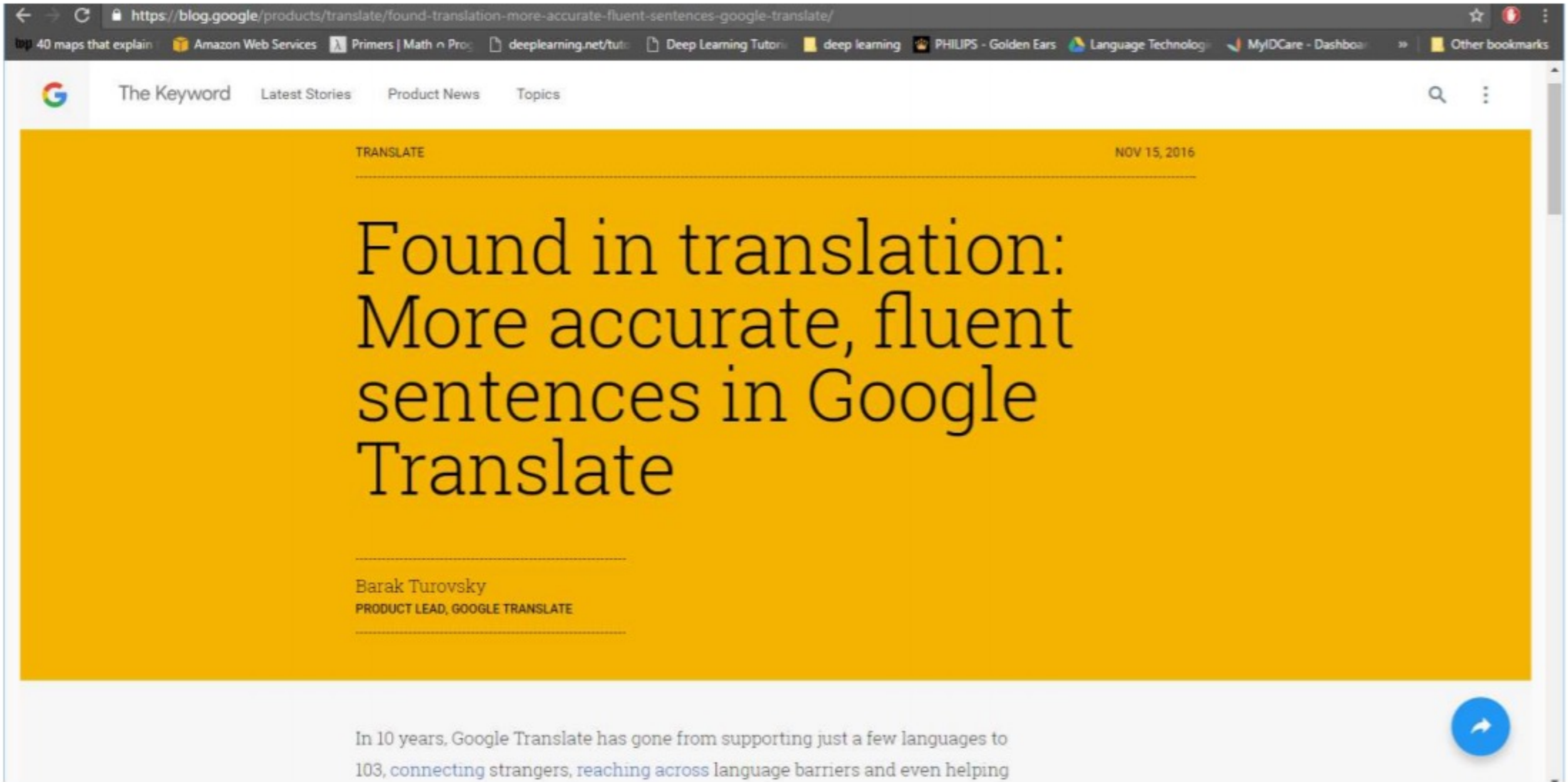
The 2018 Turing award was given to Yoshua Bengio, Geoff Hinton, and Yann LeCun.



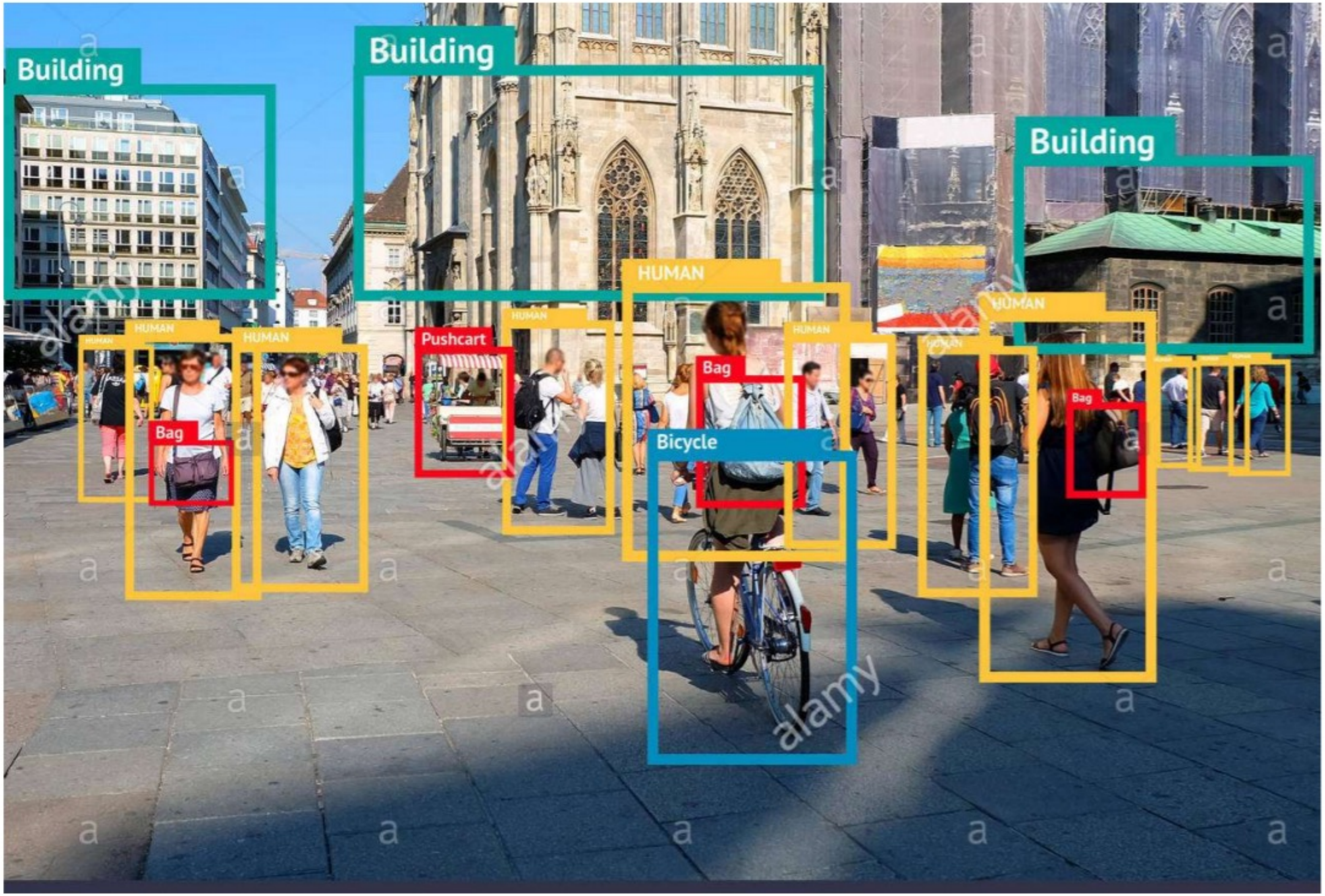
Success stories: Speech recognition



Success stories: Machine Translation



Success stories: Image segmentation



Short History of ML

- Legendre and Gauss – linear regression, 1805
  - Astronomy applications
- Probabilistic models
  - Bayes and Laplace - Bayes Theorem, 1812; Markov chains, 1913
- Fisher – linear discriminant analysis for classification, 1936
  - Logistic regression, 1940
- Rosenblatt - Perceptron, 1958
- Widrow and Hoff - ADALINE neural network, 1959
- Nelder, Wedderburn - generalized linear models, 1970
- "AI winter", limitations of perceptron and linear models, 1970
- Breiman, Friedman, Olshen, Stone - decision trees (non-linear models), 1980
- Cortes and Vapnik - SVM with kernels, 1990
- Breiman: Bagging, 1994; Ho – random forest, 1995; Freund and Shapire – AdaBoost, 1997
- Geoffrey Hinton, Deep learning, back propagation, 2006
- C. Szedegy: Adversarial manipulation of image classification, 2013

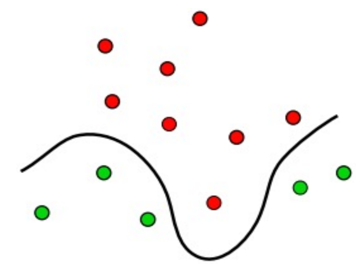
Supervised Learning

Training

Data → Feature extraction → ML model

Labeled

$x_i, y_i \in \{\text{Positive, Negative}\}$



$h(x)$

Testing
Inference

New data → Predictions

Unlabeled

x'

$y' = h(x')$

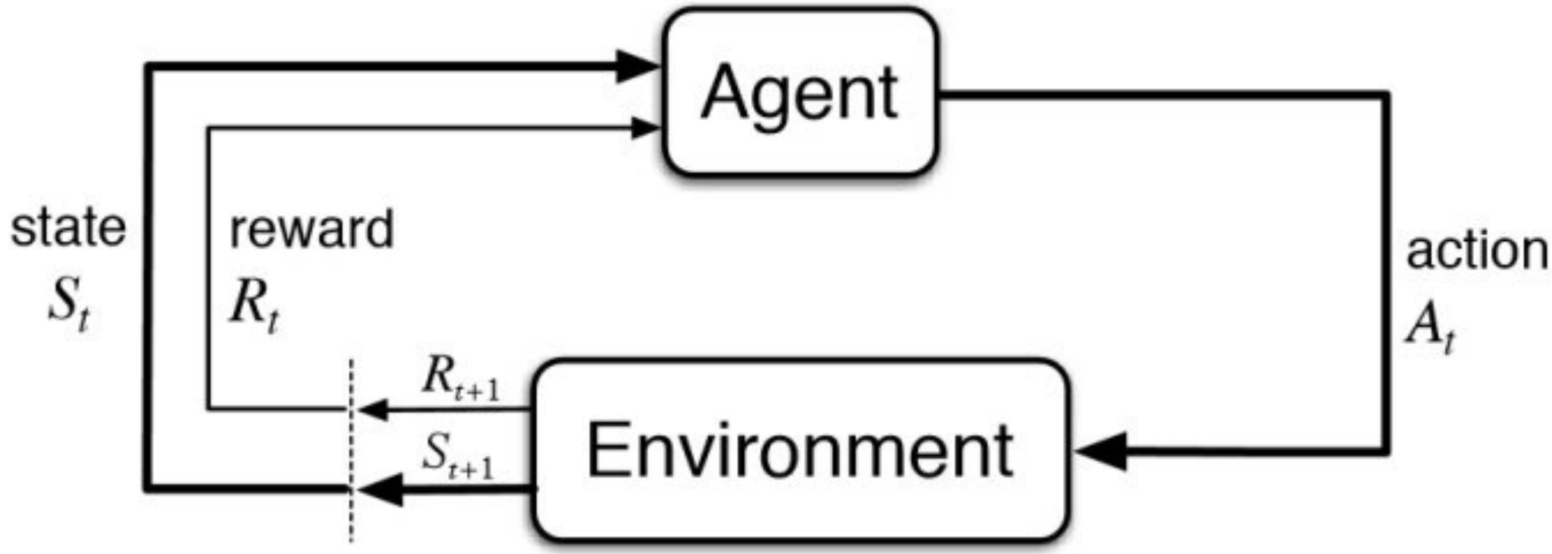
Positive
Negative

- Main Assumption: Distribution of training and testing data is similar
- Model can learn from training data and generalize to testing data
- Concrete metrics to measure model performance

Unsupervised Learning

- Input: unlabeled data
- Clustering
  - Group similar data points into clusters
  - Examples: k-means, hierarchical clustering, density-based clustering
- Dimensionality reduction
  - Project the data to lower dimensional space
  - Examples: PCA (Principal Component Analysis), UMAP
- Anomaly detection
  - Learn normal patterns during training and identify anomalies at testing
  - Examples: KDE, auto encoders, Local Outlier Factor, Isolation Forest

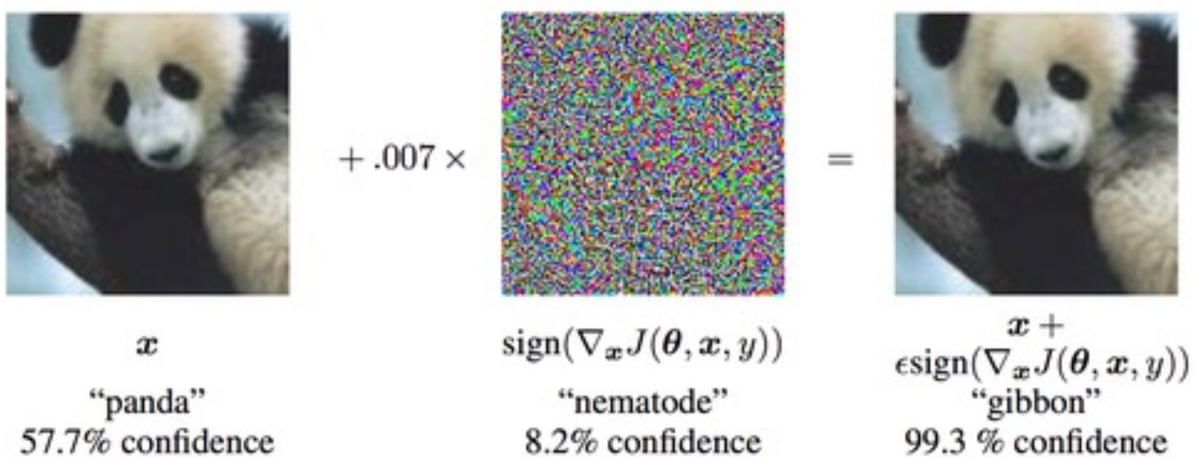
Reinforcement Learning



- Agents learn by interacting with an environment
- They take actions and obtain reward
- Goal: learn optimal policy to maximize reward
- Methods: Q learning, Deep Q Networks (DQN)
- Applications: Games (AlphaGo Zero), robotics
- https://deepmind.com/blog/article/alphago-zero-starting-scratch

Security and Privacy Risks of AI

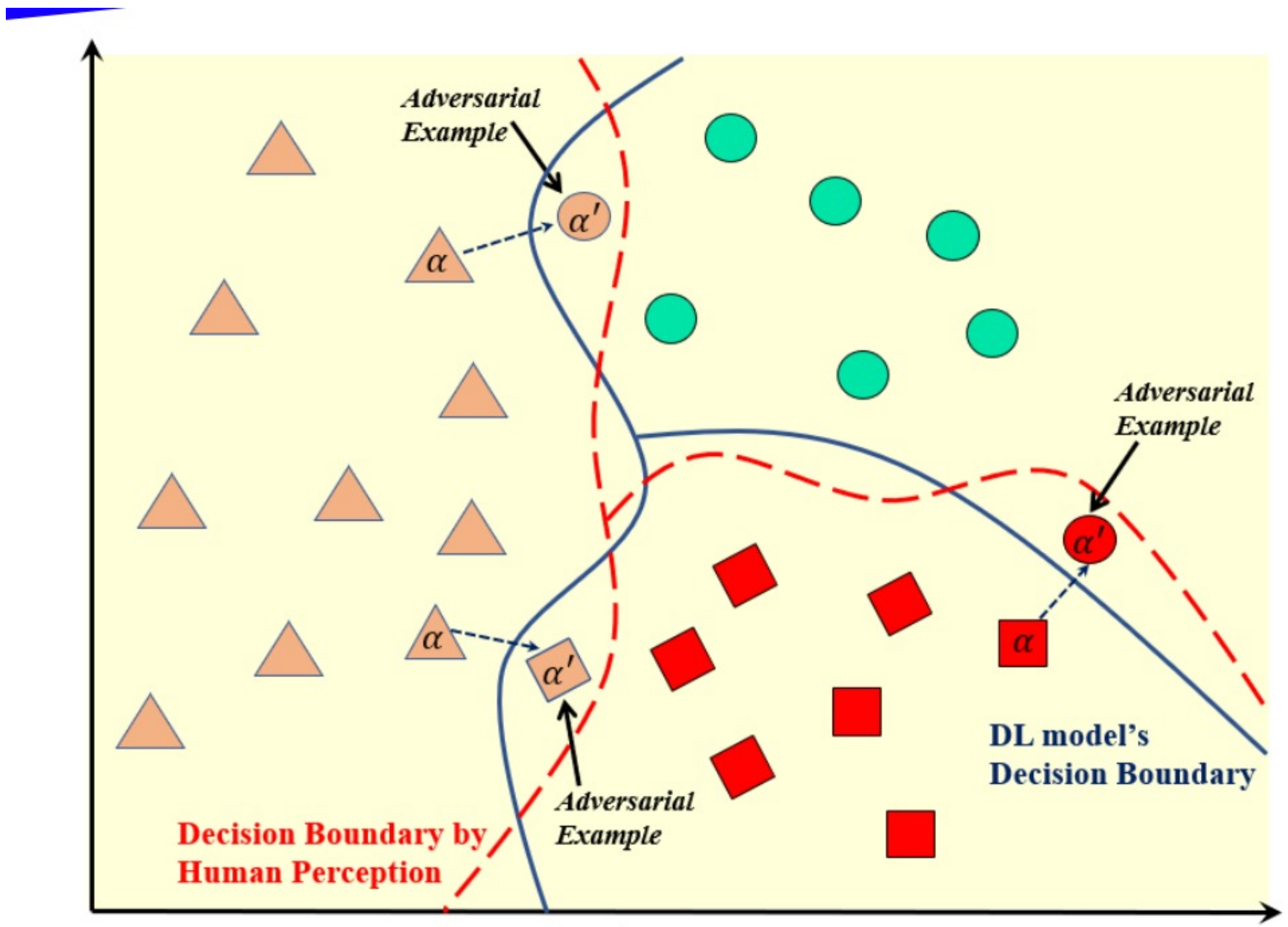
- Deep Neural Networks and other classifiers are not resilient to adversarial manipulations
  - Szegedy et al. *Intriguing properties of neural networks*. 2013
  - Biggio et al. *Evasion attacks against machine learning at test time*. 2013
  - Goodfellow et al. *Explaining and Harnessing Adversarial Examples*. 2014
- Adversarial machine learning



Attacker changes distribution of testing data!

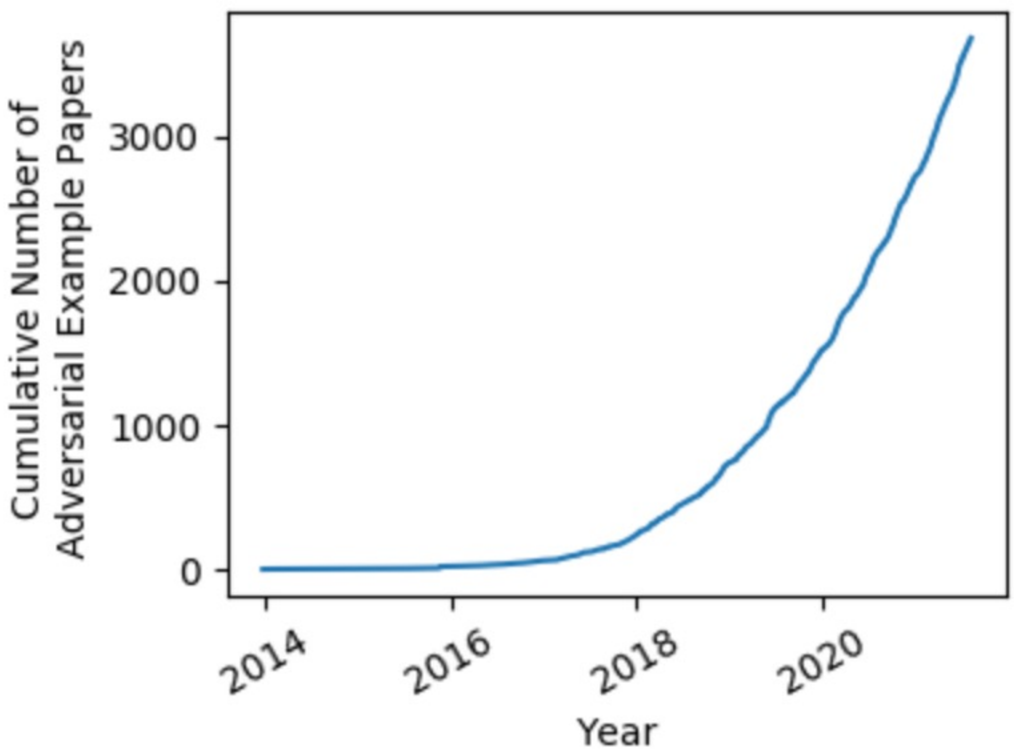
Adversarial example

What are Adversarial Examples



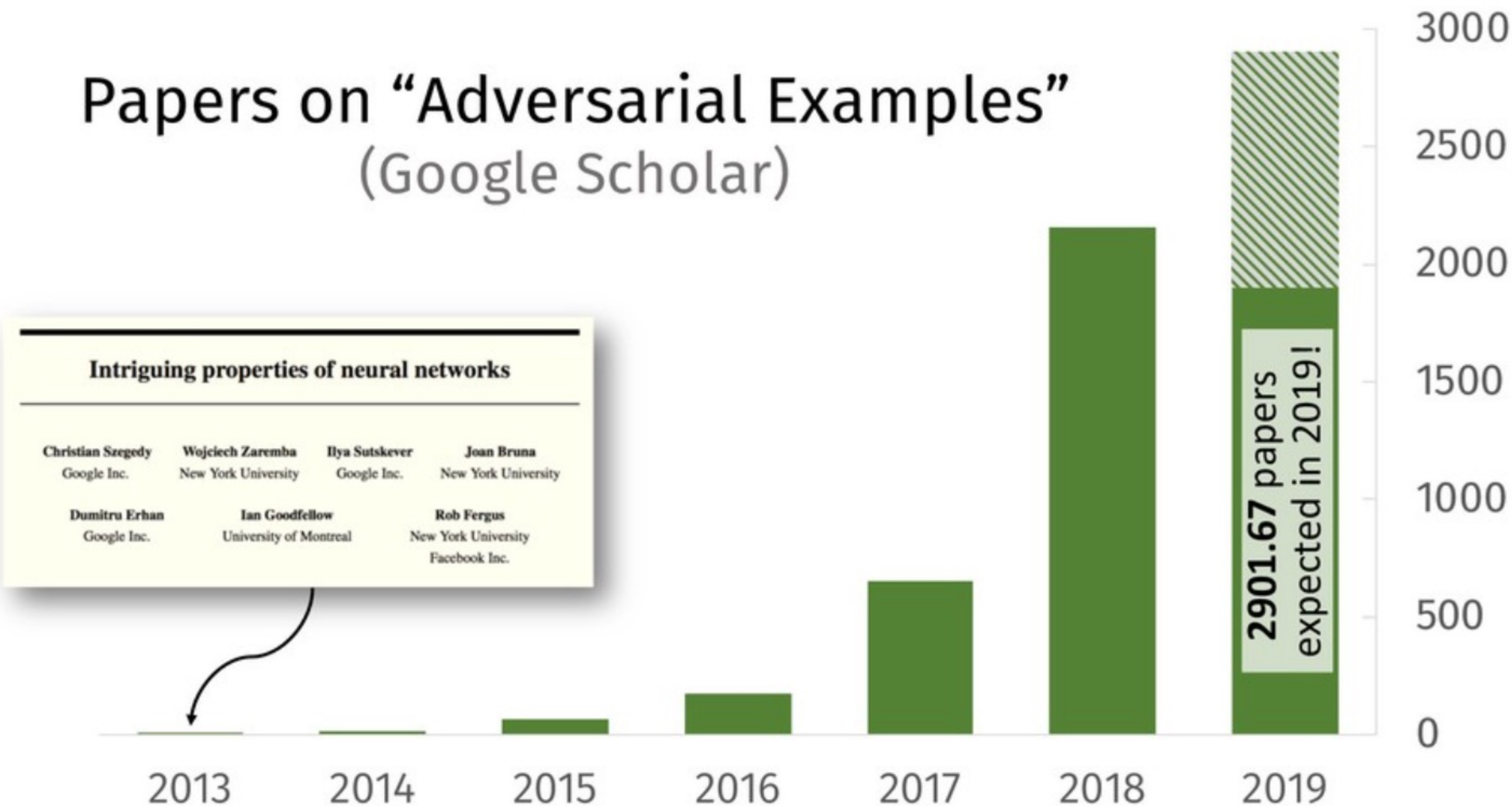
Adversarial Robustness of Deep Learning: Theory, Algorithms, and Applications. Tutorial at ICDM 2020

Adversarial ML Literature



- Graph by Nicholas Carlini, Google
- Papers published in AI and security conferences
- We will only cover a small subset (~35 papers)
- I'm always open to paper recommendations!

More Statistics



- Slide by David Evans, UVA

Safety Concerns of AI

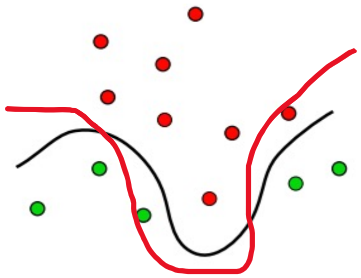
Safety Concerns of AI

- Adversarial ML
  - ML can be manipulated
  - Small change in input results in different prediction (adversarial examples / evasion attacks)
  - Corrupted training data can modify the model (poisoning attacks)
- Privacy concerns
  - User data remains private when ML models are trained on it
- Ethics and fairness of AI
  - Predictions of ML are fair for underrepresented minorities
  - Robots will not perform harmful actions

Poisoning Attacks

Training

Poisoned Data | Clean Data → Feature extraction → ML model



$x_i, y_i \in$
$\{\text{Positive, Negative}\}$

$f(x)$

Testing

New data → Predictions → Correct prediction

Subset of data
$x \in S$

Wrong prediction on points in S

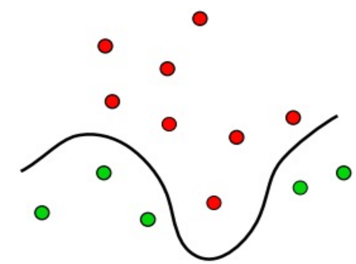
- Poisoning attack inserts corrupted data at training
- Model makes incorrect predictions on subset of data at testing

Evasion Attacks

Training

Data → Feature extraction → ML model

$x_i, y_i \in \{\text{Positive}, \text{Negative}\}$



$f(x)$

Testing

New data → Predictions

Add perturbation

$x + \Delta$

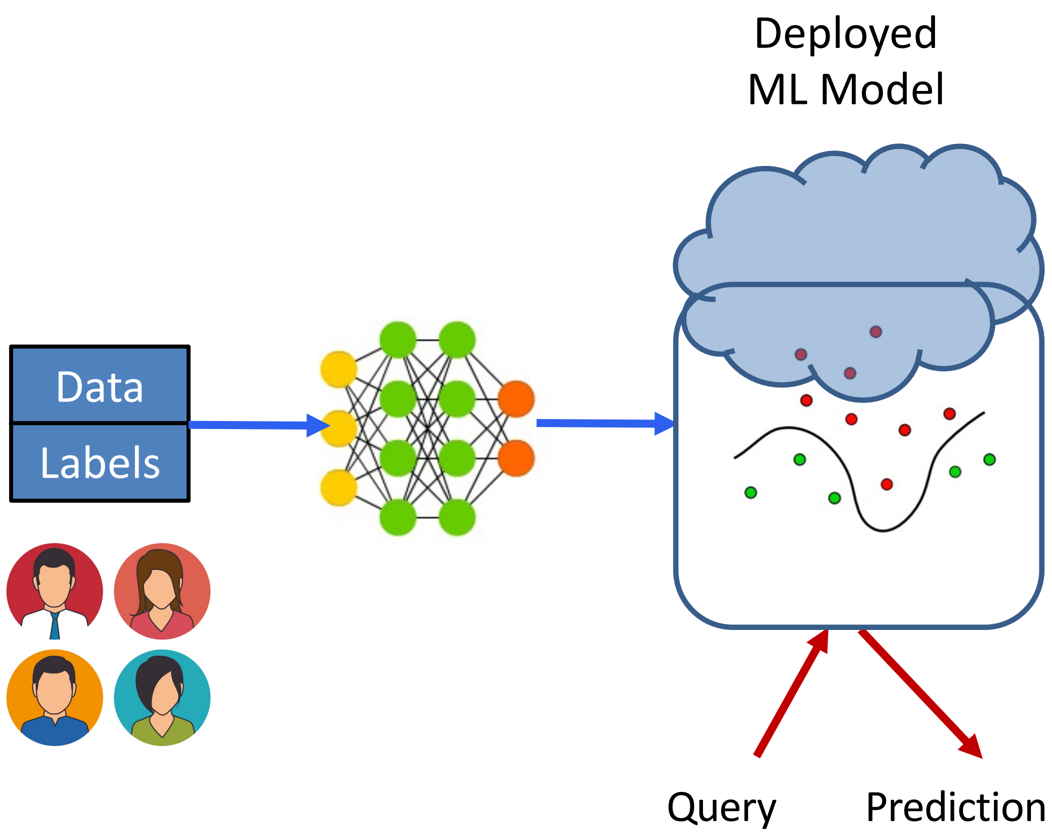
Positive
Negative

$y' = f(x + \Delta) \neq y$

Wrong prediction

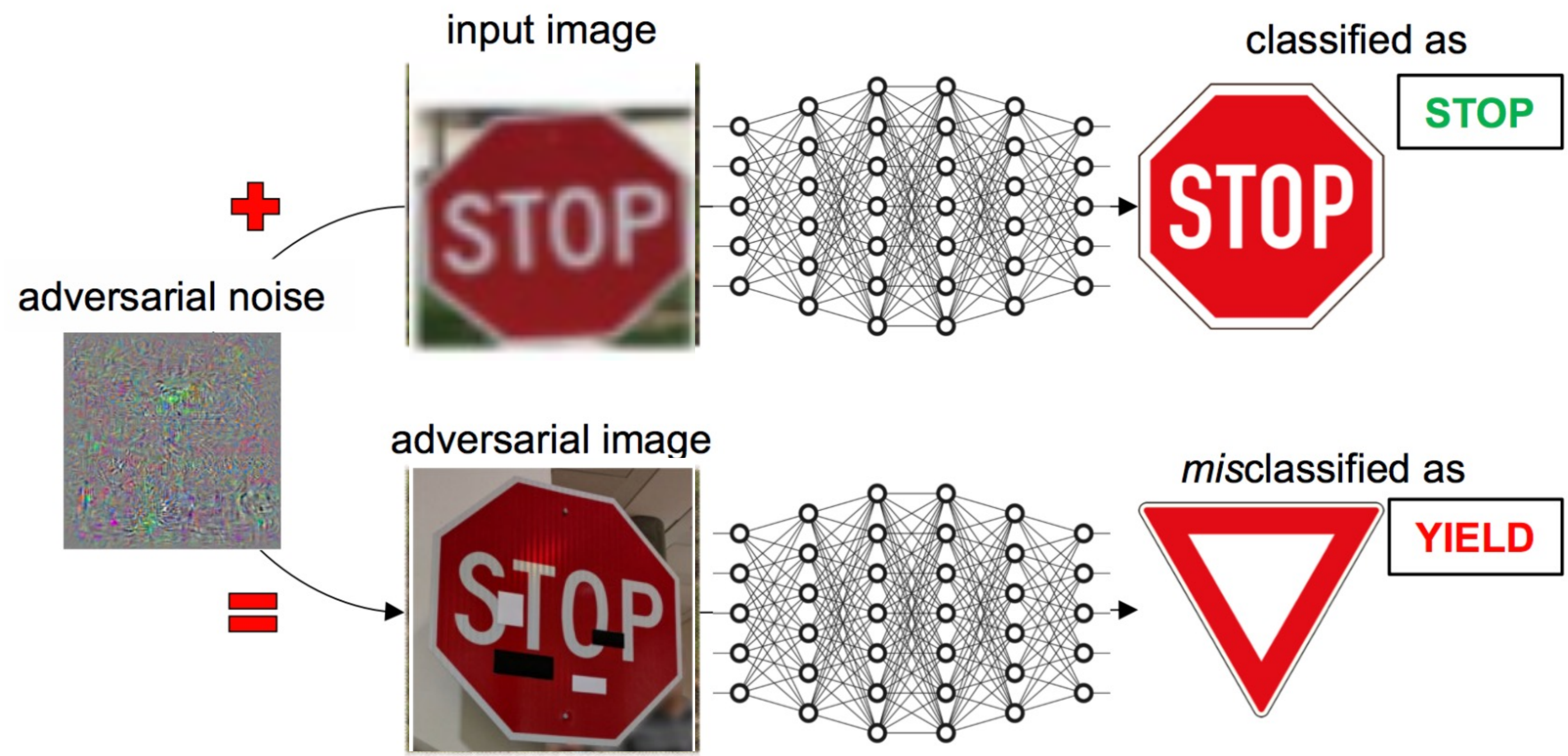
- Modify testing point by adding small perturbation to misclassify it

Privacy Attacks on ML



- Reconstruction attacks: Extract sensitive attributes
  - [Dinur and Nissim 2003]
- Membership Inference: Determine if sample was in training
  - [Shokri et al. 2017], [Yeom et al. 2018], [Hayes et al. 2019], [Jayaraman et al. 2020]
- Model Extraction: Learn model architecture and parameters
  - [Tramer et al. 2016], [Jagielski et al. 2020]
- Memorization: Extract training data from queries to the model
  - [Carlini et al. 2021]

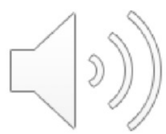
Adversarial Attacks on Road Signs



Eykholt et al. *Robust Physical-World Attacks on Deep Learning Visual Classification*. In CVPR 2018

Adversarial attacks on Speech Recognition

Audio Adversarial Examples

Audio	Transcription by Mozilla DeepSpeech
	"without the dataset the article is useless"
	"okay google browse to evil dot com"

https://nicholas.carlini.com/code/audio_adversarial_examples/

Attacking Face Recognition

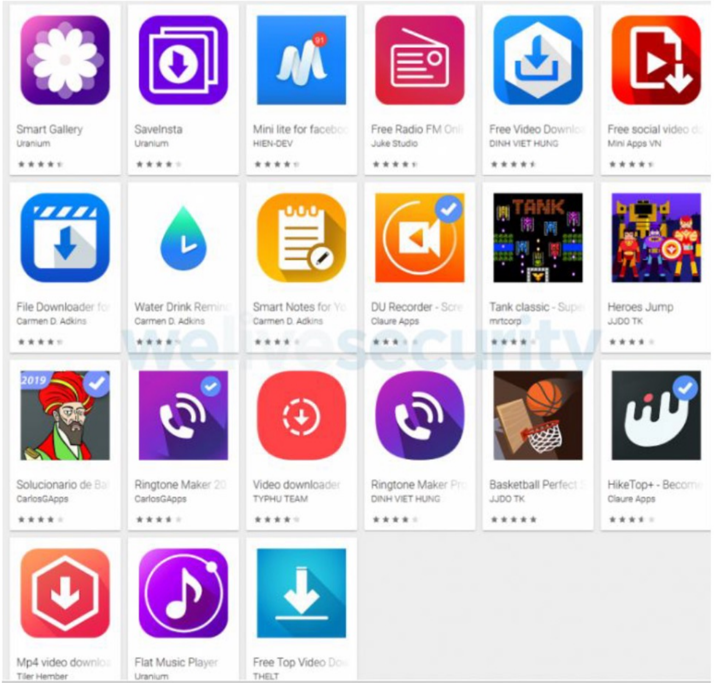
Adversarial Glasses

- M. Sharif et al. (ACM CCS 2016) attacked deep neural networks for face recognition with carefully-fabricated eyeglass frames

- When worn by a 41-year-old white male (left image), the glasses mislead the deep network into believing that the face belongs to the famous actress Milla Jovovich

Adv Attacks on Malware Detection

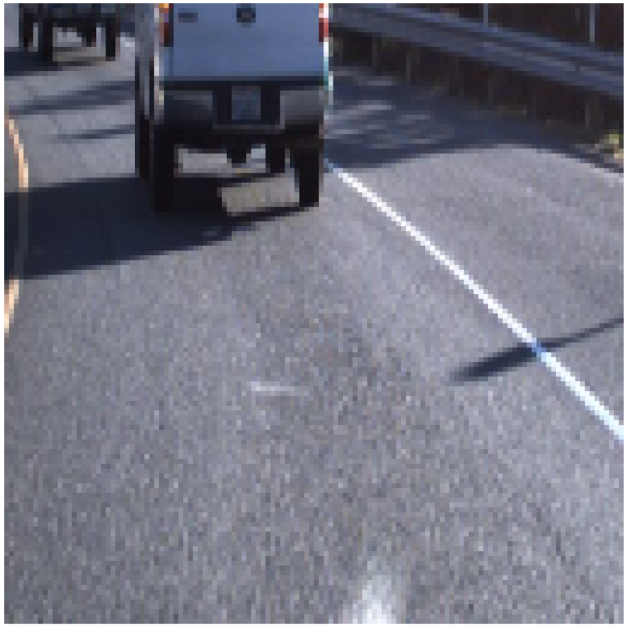
- Mislead 60% to 80% of the malicious application samples



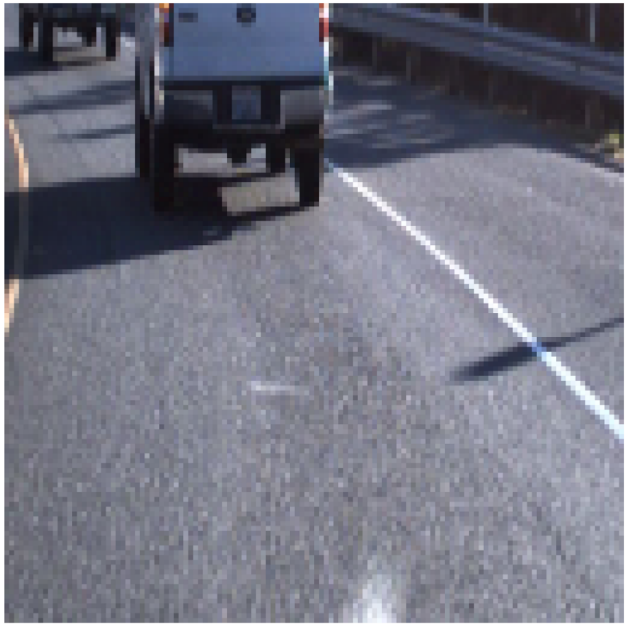
[Grosse et al, 2016](#)

Newly discovered 42 malicious apps on Google Play store [Rohit KVN, 2019](#)

Adversarial Examples in Connected Cars



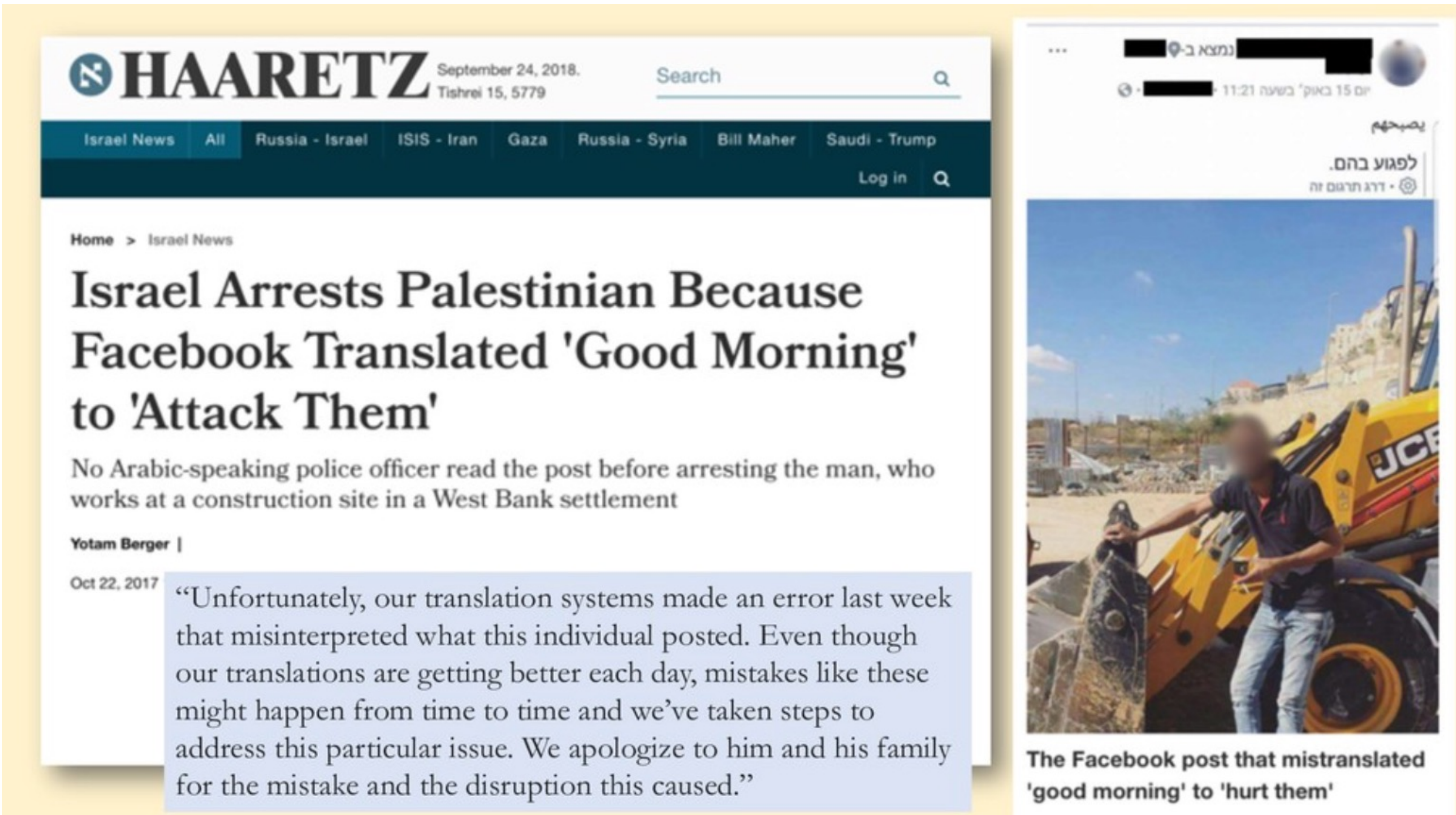
Original Image
Steering angle = -4.25



Adversarial Image
Steering angle = -2.25

- Udacity challenge: Predict steering angle from camera images, 2014
- A. Chernikova, A. Oprea, C. Nita-Rotaru, and B. Kim. *Are Self-Driving Cars Secure? Evasion Attacks against Deep Neural Networks for Self-Driving Cars*. 2019

Adversarial ML in the Real World



HAARETZ September 24, 2018. Tishrei 15, 5779

Search

Israel News | All | Russia - Israel | ISIS - Iran | Gaza | Russia - Syria | Bill Maher | Saudi - Trump | Log in

Home > Israel News

Israel Arrests Palestinian Because Facebook Translated 'Good Morning' to 'Attack Them'

No Arabic-speaking police officer read the post before arresting the man, who works at a construction site in a West Bank settlement

Yotam Berger |

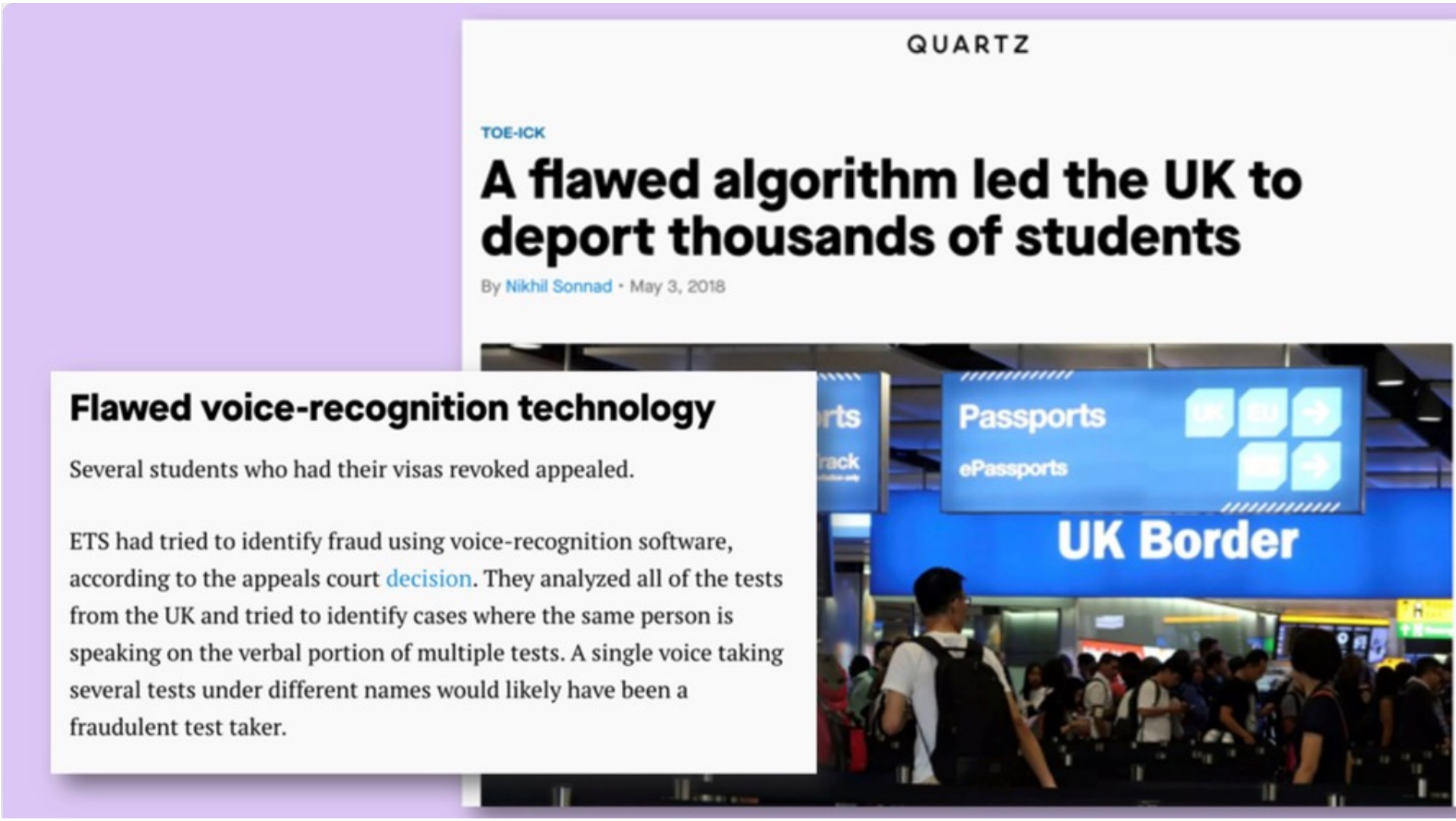
Oct 22, 2017

"Unfortunately, our translation systems made an error last week that misinterpreted what this individual posted. Even though our translations are getting better each day, mistakes like these might happen from time to time and we've taken steps to address this particular issue. We apologize to him and his family for the mistake and the disruption this caused."

The Facebook post that mistranslated 'good morning' to 'hurt them'

Slide from David Evans, UVA

Adversarial ML in the Real World



QUARTZ

TOE-ICK

A flawed algorithm led the UK to deport thousands of students

By Nikhil Sonnad • May 3, 2018

Flawed voice-recognition technology

Several students who had their visas revoked appealed.

ETS had tried to identify fraud using voice-recognition software, according to the appeals court decision. They analyzed all of the tests from the UK and tried to identify cases where the same person is speaking on the verbal portion of multiple tests. A single voice taking several tests under different names would likely have been a fraudulent test taker.

Slide from David Evans, UVA

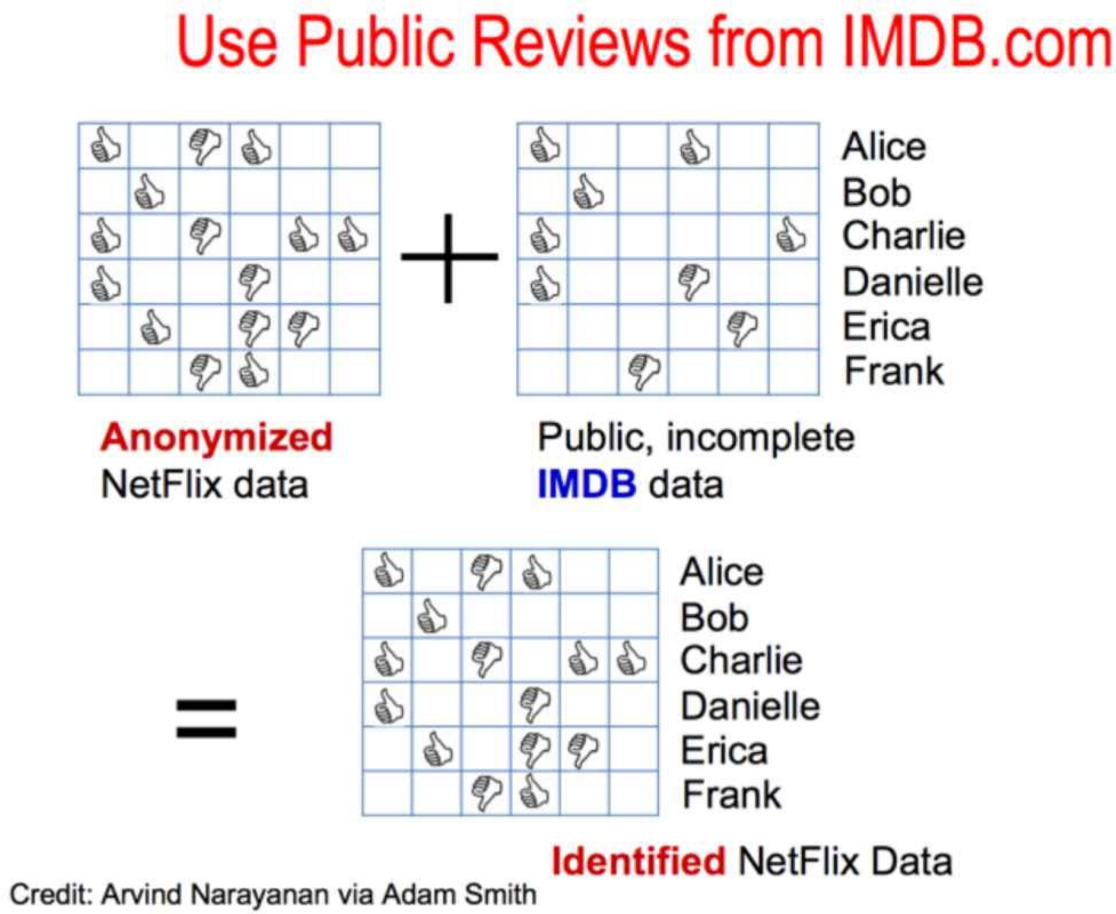
Poisoning in the Real World

Listen to this article

It took less than 24 hours for Twitter to corrupt an innocent AI chatbot. Yesterday, Microsoft unveiled Tay — a Twitter bot that the company described as an experiment in "conversational understanding." The more you chat with Tay, said Microsoft, the smarter it gets, learning to engage people through "casual and playful conversation."

Privacy Attacks

Deanonymizing Netflix Data



Narayanan, Shmatikov, [Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)](), 2008

Summary

- AI has a long history
- Adversarial ML gained attention with the discovery of adversarial examples by Szedegy et al. 2013 and Biggio et al. 2013
- Different types of adversarial attacks
  - Poisoning (training time)
  - Evasion (inference time)
  - Privacy
  - Fairness
- Multiple application domains: image classification, speech recognition, cyber security
- Defenses are usually domain specific and not fully working